Article

Patterns

Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations

Highlights

- Scholexplorer data can be used to identify reuse and citation of published datasets
- More dataset and article links can be identified now with the Scholexplorer API
- Many links result from former manual data curation instead of direct data citation
- Author and dataset owner affiliation would help identify different data use cases

Authors

Nushrat Khan, Catherine J. Pink, Mike Thelwall

Correspondence

n.j.khan@bath.ac.uk

In Brief

Identifying links between articles and supporting data is vital for demonstrating reuse and impact of published data. Scholix creates these links, and we find that the Scholexplorer API can locate more article-dataset links than was previously possible in practice. Our study finds evidence of data reuse, but we suggest that further enhancement of the Scholix schema and enrichment of Scholexplorer metadata through controlled vocabulary and inclusion of persistent identifiers would recover more cases of secondary data use.

Khan et al., 2020, Patterns *1*, 100007
April 10, 2020
https://doi.org/10.1016/j.patter.2020.100007

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations

Nushrat Khan,[1,2,3,*] Catherine J. Pink,[1] and Mike Thelwall[2]
[1]Library Research Data Service, University of Bath, Bath, Somerset BA2 7AY, UK
[2]Statistical Cybermetrics Research Group, University of Wolverhampton, Wolverhampton WV1 1LY, UK
[3]Lead Contact
*Correspondence: n.j.khan@bath.ac.uk
https://doi.org/10.1016/j.patter.2020.100007

THE BIGGER PICTURE The number of research data repositories has substantially increased in response to growing requirements for publication of data supporting research findings. However, the lack of a common language between repositories and journals makes it difficult to find connections between datasets and articles and to identify secondary data-reuse cases. This study explores how the Scholix (Scholarly Link eXchange) framework can be used to create these links in order to validate research findings, to demonstrate compliance with funder mandates, and to understand the value and impact of research data. This is the first quantitative analysis of data gathered from the Scholexplorer API and demonstrates its potential for identifying data reuse. A content analysis of citing articles reusing data also shows that few of these links resulted from standard data citation practice. The findings of this study provide the basis for further comparative analyses to develop standard community practices.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

The Scholexplorer API, based on the Scholix (Scholarly Link eXchange) framework, aims to identify links between articles and supporting data. This quantitative case study demonstrates that the API vastly expanded the number of datasets previously known to be affiliated with University of Bath outputs, allowing improved monitoring of compliance with funder mandates by identifying peer-reviewed articles linked to at least one unique dataset. Availability of author names for research outputs increased from 2.4% to 89.2%, which enabled identification of ten articles reusing non-Bath-affiliated datasets published in external repositories in the first phase, giving valuable evidence of data reuse and impact for data producers. Of these, only three were formally cited in the references. Further enhancement of the Scholix schema and enrichment of Scholexplorer metadata using controlled vocabularies would be beneficial. The adoption of standardized data citations by journals will be critical to creating links in a more systematic manner.

INTRODUCTION

In recent years there has been a major push from funders to make research outputs, including research datasets, openly available.[1] As a result, there are increasingly many research data infrastructures within higher education institutions and policies for research data support. Organizations and committees, such as the Research Data Alliance, FORCE11, and CODATA (the Committee on Data for Science and Technology), are supporting this rapidly changing research environment and tackling emerging issues in the field of research data management. However, due to differences in domain practices and requirements from the funders, different fields have been moving at their own pace to adopt and adapt cultures of data sharing.

Due to the substantial time and effort required to document and share high-quality research data, it is important to know whether shared data will be reused.[2,3] Several studies have explored associations between shared research data and the citation rates of articles in different fields, such as cancer microarray data in genomics,[4] astronomy,[5] astrophysics,[6] and for open-access journal articles published from PLOS and BMC.[7] All report higher citation impact for articles sharing research data.

Despite the evidence of positive citation impact for articles that share research data, fewer studies have explored data citation practices and reuses of shared data. This is largely due to the lack of standards in data citation practice across different fields and journals. Mayo et al.[8] investigated data citation

practices for articles in the Dryad repository and found that the number of articles that cite data in their citation sections increased from 5% to 8% between 2011 and 2014 while intratextual citation had grown from 69% to 83%. Khan et al.[9] report that 27% of articles citing biodiversity datasets indexed by the Global Biodiversity Information Facility cited them in the methods and references, 13% in the methods and data access statements, and the rest (58%) intratextually and in supplementary information, with only 2% mentioned in all three sections. Articles using a large number of datasets (12–50) cited them inconsistently. Colavizza et al.[10] investigated 531,889 journal articles published by PLOS and BMC and found that "following mandated publisher policies, data availability statements have become common by now, yet statements containing a link to a repository are still just a fraction of the total." This is possibly why an attempt to capture data citation using Thomson Reuters' Data Citation Index (DCI) by Robinson-García et al.[11] found that 88.1% records received no citations because DCI harvests citations that are cited in a standard format in the references. Mathiak and Boland[12] and Ghavimi et al.[13] explored variations in citation practices for social sciences datasets, and the latter study proposed a linked data approach to solve this issue by developing an ontology.

To resolve these issues, Silvello[14] suggests that "… [T]he ideal data citation system should uniquely identify a dataset and subsets of it with different levels of coarseness (identification), attribute the ownership and responsibility of the data with variable granularity to the right people/institutions (attribution), guarantee the persistence of the data being cited as well as the citations themselves (fixity), and automatically create complete and consistent citation snippets (completeness and consistency) according to community practices and shared metadata standards."

While more journal publishers are adopting standardized ways to link research datasets by including data access statements or similar sections, we are still far from reaching an agreement on how data should be cited. This has further complicated the task of finding links between articles and datasets.

When investigating article-dataset links, two types of article and dataset relationships are considered important by the data producers (e.g. researchers, doctoral students), institutional repository managers (e.g. data librarians, archive support staff), and research managers (e.g. pro-vice-chancellors, institute directors, and research committees): (1) links between primary datasets and research articles associated with them to prove compliance with funder mandates; and (2) links between a dataset and any articles that reused it, demonstrating the impact of that dataset. Nevertheless, most research data repositories currently act as silos and searches by author affiliation are often not viable. Thus, finding datasets published by institutional researchers and data producers in external archives can be arduous and not feasible. The problem is particularly acute for institutional repository managers, who must rely on web searches or manual notifications from data producers when they publish their data in an external repository in order to be able to report on policy compliance or impact to research managers. Moreover, inconsistency in citation practices, as mentioned above, makes it difficult to find evidence of data reuse and citation by secondary data users. As a result, data producers may be unable to demonstrate the impact of their published data or claim full credit for their work. This leaves a big knowledge gap for both groups, and especially for institutional repository managers who need to maximize their resources by developing systematic approaches to identifying article and dataset links in external repositories. It is important to fill this gap to check whether data producers are complying with funder and publisher requirements to make their data openly available for both reproducibility and reuse.

The recently developed Scholix (Scholarly Link eXchange) framework is based on establishing links between datasets and articles using event data published by DataCite and Crossref.[15] Data collected using this framework is aggregated by Scholexplorer and made freely available by its REST application programming interface (API).[16,17] Multiple articles have discussed the mechanisms and scholarly benefits of this framework.[7,18,19] Limani et al.[20] used an alternative approach to establish links between research datasets in the Journal Data Archive and publications about the economy that were published in the EconBiz portal, and reported that the links found using their approach could be valuable for Scholix. Several higher educational institutions such as the University of Manchester,[21] Durham University,[22] and the University of Illinois at Urbana-Champaign[23] have explored the Scholexplorer data and developed individual processes to incorporate it into their systems. However, no published empirical studies have analyzed output data from the Scholexplorer API to identify: (1) who publishes the linked datasets; (2) whether the data can identify data-reuse cases; and (3) how typical links are generated.

To explore the usability and quality of the data derived from Scholexplorer, this article built on and extended Python code originally developed by Durham University[22] to collect data from the Scholexplorer API for approximately 31,890 research outputs published by University of Bath researchers. This includes all research outputs recorded in the university's CRIS (current research information system) and publications repository, Pure, until April 17, 2019. University of Bath systems were used as the data source for this case study to derive a comprehensive and reliable list of research output digital object identifiers (DOIs) from Pure. This was required because the Scholexplorer API does not yet support affiliation search, and to be able to use the University of Bath's Research Data Archive (UBRDA) as a benchmark to compare against the Scholexplorer API output.

By November 2019, there were 332 UBRDA datasets registered on DataCite.[24] Staff in the University of Bath Library's Research Data Service who support the UBRDA have developed methods to locate datasets published by the university researchers in some external archives. However, these methods are too resource intensive for a small team to routinely use, which is likely to be a common problem internationally that is time consuming to address. However, as an aggregator of data from many journal and data publishers, Scholexplorer might provide a systematic approach for solving this issue. The research questions of this case study assess this potential from the perspective of institutional repository managers.

RQ1. Can the Scholexplorer API identify previously unknown links between university research outputs and datasets in external archives?

RQ2. Can Scholexplorer identify examples of data reuse?

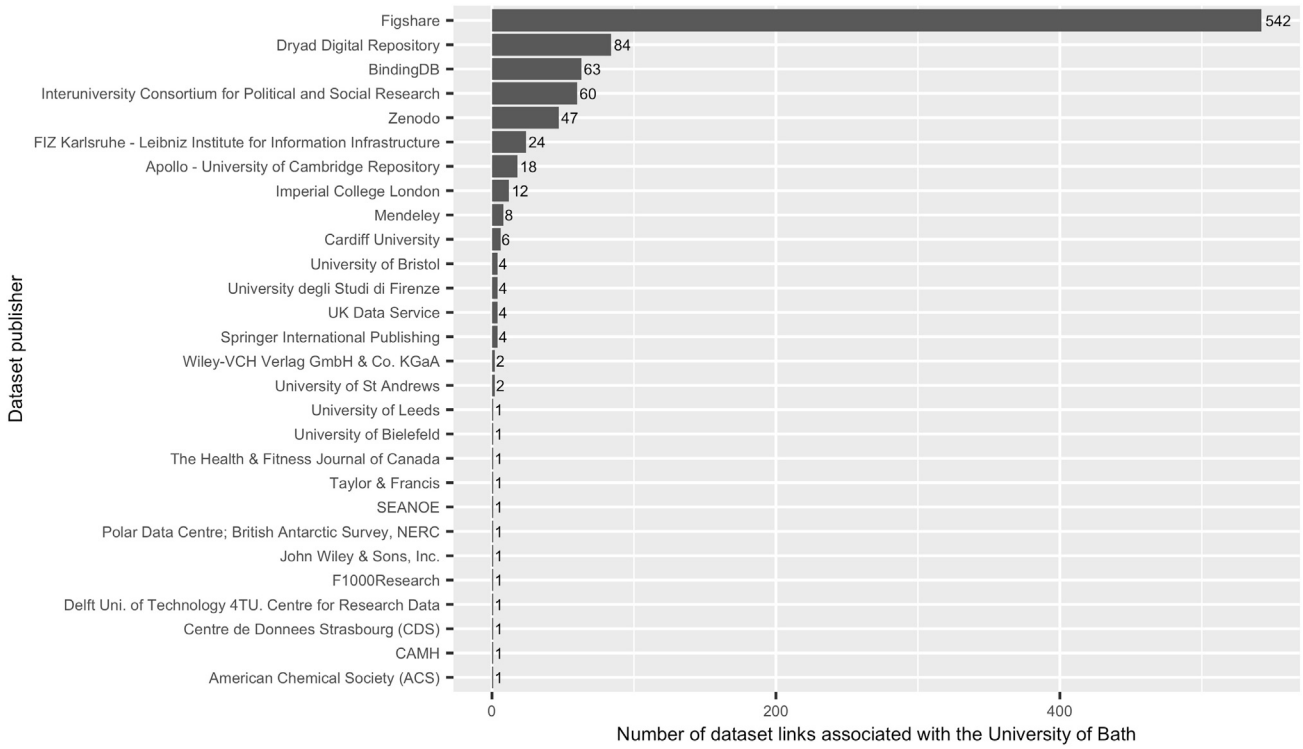

Figure 1. Dataset Links Found by Scholexplorer in External Repositories Associated with University of Bath Research Outputs, Excluding CCDC

RESULTS

The Scholexplorer API was tested for University of Bath research outputs to investigate whether it provided useful information, as described in Experimental Procedures.

RQ1. Scholexplorer Identified Previously Unknown Dataset-Article Links

As of September 2019, UBRDA recorded 48 University of Bath-affiliated datasets that were published in external repositories. These had either been reported by the researchers themselves or had been identified by data librarians managing the UBRDA and manually searching for missing connections. The Scholexplorer API identified 1,501 unique research outputs with at least one University of Bath author linked to at least one dataset, a 31-fold increase. In total 5,002 datasets were associated with these 1,501 research outputs, where one output is linked to one or more datasets. Most of the datasets were from the Cambridge Crystallographic Data Center (CCDC) (82.1%). This is in line with the findings by Robinson-García et al.[11] that crystallography accumulated more than half of all citations to datasets on DCI. However, it can be difficult to find those links using the single search system on the CCDC (https://www.ccdc.cam.ac.uk/structures/) where the search functionality is limited to identifiers, compound names, DOIs, authors, journals, and publication details (year, volume, page). Advanced searching requires registration and a license, creating a barrier to simplified access. Besides CCDC, Scholexplorer also identified 28 other external repositories hosting datasets associated with University of Bath researchers (Figure 1).

The Scholexplorer API identified four datasets on the UK Data Service (UKDS) that had linked to two journal articles affiliated with University of Bath researchers. The API has some gaps in coverage, however. UBRDA had identified seven dataset records associated with University of Bath researchers published on the UKDS for Economic and Social Research Council-funded projects, but none were recovered through our use of the Scholexplorer API (Table 1). One of the datasets (10.5255/UKDA-SN-852040) had associated journal article links on its UKDS record that had not been identified by Scholexplorer. This could be because this dataset is not part of the new beta UKDS website but is on the separate UKDS ReShare repository, which is increasingly used by researchers to self-archive datasets from "long-tail" research studies. It is possible that metadata from the ReShare domain may not yet have been consumed by the Scholexplorer API. Datasets in UKDS ReShare have a unique ID ending with UKDA-SN-(6-digit number).

For the other six missing datasets, UKDS did not link to any related journal articles. However, links to reports, a book chapter, and working papers appear on project websites linked to these dataset records (10.5255/UKDA-SN-8303-1, 10.5255/UKDA-SN-8176-1, and 10.5255/UKDA-SN-8397-1). Perhaps because those publications were linked using general URLs instead of DOIs, Scholexplorer has not indexed them. Persistent identifiers for reports and white papers would therefore help with capturing the value of any datasets used, although this is not currently common practice in many organizations. Similarly,

Table 1. Links to UKDS Datasets Associated with Bath Researchers

DOIs of UKDS Datasets Listed on UBRDA	DOIs of UKDS Datasets Identified from Scholix
10.5255/UKDA-SN-8397-1	10.5255/UKDA-SN-5050-16
10.5255/UKDA-SN-8176-1	10.5255/UKDA-SN-7480-1
10.5255/UKDA-SN-8303-1	10.5255/UKDA-SN-7649-1
10.5255/UKDA-SN-852527	10.5255/UKDA-SN-7260-1
10.5255/UKDA-SN-852064, 10.5255/UKDA-SN-852065 (Datasets: parts 1 and 2 from the same project)	
10.5255/UKDA-SN-852040	

searches of the UK Research and Innovation website for the datasets 10.5255/UKDA-SN-852527, 10.5255/UKDA-SN-852064, and 10.5255/UKDA-SN-852065 found associated journal articles that were not linked to the datasets on UKDS. As for the dataset 10.5255/UKDA-SN-852040 mentioned above, these links were not indexed by the Scholexplorer API. Thus, while its coverage is gradually increasing, output from the Scholexplorer API is not yet comprehensive and some of the gaps seem to be inevitable.

In the first phase of data collection from the Scholexplorer API in September 2019, only 41 journal articles out of 1,501 (2.7%) included research articles' author names in their Scholexplorer API records. These 41 journal articles were further investigated to identify whether the associated datasets were primary (by the same authors) or secondary (different authors). For the 121 related datasets (one publication can be linked to more than one dataset) there were 10 cases of dataset reuse. Most were from the social sciences and published by the Inter-university Consortium for Political and Social Research (ICPSR) (n = 7). The rest were from the UKDS (n = 1), FigShare (n = 1), and Binding DB (n = 1). A further analysis of these citing articles conducted to validate proper data citation practice is reported below.

RQ2. Scholexplorer Can Be Used to Identify Reuse of Published Data

The Scholix schema is based on a simple source-and-target relationship in which the links come from a provider (dataset publisher or journal article publisher). Metadata supported for target and source include identifier, object type, title, creator, publication date, and publisher, where only identifier and object type are compulsory and the rest are optional fields (0 … N). This limited metadata availability allows simplicity but can leave gaps. For example, author names were only available for 2.7% of the research articles in the September dataset.

As of June 2018, there were more than 870,000 links between Crossref DOIs and DataCite DOIs aggregated by the Scholexplorer API, where most of the links originated from DataCite DOIs and only 22,000 links from Crossref journals. Of those 22,000 Crossref DOIs, only 16% (3,657) were links between dataset and literature defined by using a Crossref type for a scholarly text document and the DataCite metadata resourceTypeGeneral of "Dataset," and the rest could be literature-literature links (where both source and target objects in the API's JSON output are literature).[25] It is not clear whether the author information was missing mainly from the original publisher information supplied to Crossref. These findings were communicated to the Scholix group members (Adrian Burton, Martin Fenner, Wouter Haak, and Paolo Manghi) by an email on October 07, 2019. A second set of metadata was then collected from the Scholix API for the 1,501 research output DOIs extracted in the first phase that had been linked to at least one unique dataset, with a substantial increase (from 2.7% to 89.2%) in the number of author names available for research articles linked to datasets, perhaps due to updates in the Scholexplorer API software. This suggests that the quality of data is improving and that the Scholix group is responsive to user feedback, which represent positive indicators of the likely continued value of the framework.

By manually comparing author names of articles and datasets of the 41 journal articles and 121 associated datasets from the initial data collection, as mentioned above, we identified ten studies with evidence of data reuse. To explore whether these examples of data reuse identified through the Scholexplorer API had employed standard data citation, we examined the citing articles. Most of the datasets (7 out of 10) were not included in the reference section of the articles, six studies reused data from ICPSR and mentioned it in the methods section, and two cited the associated survey websites but not the datasets. Given that we are not aware of Crossref text mining full text of articles for references, we conclude that the datasets in ICPSR were likely to have been linked manually by a data curator or staff at ICPSR. While it is useful to learn about such data-reuse cases, it does not demonstrate that links are commonly established due to research articles citing datasets directly in the article references.

One article (10.1016/j.chembiol.2010.07.018) had links to two BindingDB datasets (10.7270/q2jd4v85 and 10.7270/q2n014t9), which were not cited in the article, but the article was linked on BindingDB records. By comparing the dataset records and author names and affiliations, we identified the first dataset to be the primary research output and the latter as a case of secondary data reuse. Even though BindingDB links associated articles with the datasets on its platform, dataset creator names are not included in the Scholexplorer records. Both the creator and publisher metadata fields include "BindingDB" only, which is not useful for automating identification of data-reuse cases. The rest of the articles (n = 3) reused datasets from FigShare, ICPSR, and UKDS and had included citation for both primary (n = 2) and secondary datasets. Among the four UKDS datasets identified by Scholexplorer, three were cited by one journal article (DOI: 10.1186/s12889-017-4665-1) with citations included in the reference section.

The Python code used to create the study datasets using the Scholexplorer API for the same set of 1,501 research output DOIs found fewer records during the second analysis: 5,079 in September 2019 compared with only 5,002 in November 2019. We assume this to be a technical issue, as we could not identify any other reason. This should be considered when downloading and using data from the Scholexplorer API for creating metadata records to external datasets in an institutional archive because maintaining consistency is integral to this process.

Variations in repository names can also cause problems. For example, Dryad and Dryad Digital Repository, FigShare and

FigShare Academic Research System, and Zenono and Zenodo Research Shared are all different variations of three repositories that had to be merged when cleaning the dataset in order to calculate the total number of links found in each repository for Figure 1. Most of the data repositories, including Dryad, FigShare, and Zenodo, are members of DataCite, and according to the Scholix website[15] the method of participation is to feed the data-literature link information to DataCite, which is then aggregated by Scholexplorer. As noted by Aaron Tay in his blog post,[23] it is not clear how the Scholix schema is implemented in every repository, how those relationships are mapped, or whether this affects the API output. Better documentation of metadata mapping and further analysis of the API data would therefore be welcome in order to develop a benchmark. A controlled vocabulary or use of persistent identifiers could help in the future to aggregate the results efficiently. In case multiple name repository variations need to be captured, either the schema should be extended accordingly, for example using a metadata field such as isSameAs, or integration of persistent identifiers for the host organizations, such as the Research Organization Registry[26] or the repository DOI issued by re3data.org, would be necessary. Building such persistent identifier (PID)-based relationships is the focus of the European Union-funded FREYA project,[27] and incorporating these developments in the Scholix framework would help with repository name disambiguation.

DISCUSSION

This case study demonstrates that the Scholexplorer API is able to find links between articles and research datasets published in external archives that data librarians would previously have struggled to find. This article also introduces Python code that can be reused by other institutions for this purpose.[28] This study demonstrated that the information gathered from the Scholexplorer API can be used to validate the impact of the research data published by the data producers and provide evidence of compliance to funders' mandates. When datasets are deposited in a data repository external to the data producer's host institution, it can help gather information on research collaborators and generate network graphs.

Besides exposing links to related datasets for articles (e.g., Scopus, the bento-box search system from the University of Illinois at Urbana-Champaign[23]), another common use case of the Scholexplorer API by the repository managers at higher education institutions is to create metadata records with data derived from Scholexplorer links to demonstrate the impact of their researchers' datasets.[21,22] Differentiating between primary research data and secondary data-reuse cases is therefore important to give an accurate picture of impact, since the latter may not be a scholarly output from the researchers affiliated with that institution. Such connections are currently not straightforward in identifying from Scholexplorer because of the unavailability of author affiliation, missing author names in some cases (e.g., records from BindingDB), the lack of a standard naming format (e.g., initials or full first name, order of first and last name), the lack of use of author identifiers such as ORCID (Open Researcher and Contributor ID), multiple occurrences of same author names that can potentially slow down the processing speed of computer programs designed to compare them, and coverage being limited to peer-reviewed articles only.

The inclusion of author affiliation information in the Scholexplorer API metadata and addition of a search function by affiliation would greatly benefit the repositories to search by their respective institutions. At present it is only possible to search by publisher names on the Scholexplorer API, which will only return the datasets published through UBRDA and not any University of Bath-affiliated datasets published in external repositories. Currently there is an identifier field for authors in the Scholix schema that has not yet been implemented in practice. Furthermore, names alone cannot be used accurately to compare and identify a person when different naming formats are used. For example, Fear[29] took a similar approach to automate data-reuse studies by comparing author names of datasets published on ICPSR and associated articles but came to the conclusion that this may lead to erroneous results due to lack of other contextual information.

The integration of other PID services, such as ORCID,[30] would not only help to create more diverse PID relationships but also encourage researchers to adopt such services. Given that ORCID is not mandatory, this type of added value could be an incentive to promote its use. The FREYA project, with similar partners to Scholix (e.g., DataCite and Crossref), planned to integrate PID services to generate meaningful PID graphs,[27] and the results will hopefully transfer to the Scholix schema as both services mature.

The ideal way of creating article and dataset links would be to ensure that the associated dataset is linked to an article when it is published online. However, data citation has only recently started to becomecommon practice and can vary greatly in different fields. Many of the article and dataset links that are currently aggregated by Scholexplorer are the result of the manual labor of data curators. For example, ICPSR had started creating a bibliography of articles citing their datasets and even though a rich source of information, these links are not proof of improving citation practices in journals. Collaborations among journal publishers and repository managers are therefore integral to further improve the data quality of Scholexplorer and ease the process of systematic linking between articles and datasets.[31] Furthermore, our results show that gray literature, such as reports, book chapters, and working papers, are currently not covered by the Scholexplorer API. More studies with larger sample sizes should be conducted in this area to explore its coverage of gray literature, and the research community should identify how the scope of Scholexplorer can be expanded in future to address this issue, as this can be a valuable tool to identify more reuse cases and societal impact.

Finally, the Scholix project addresses an aspect of open science. The best way to promote and support this system is to openly share implementation and integration methods by different institutions for different repository platforms, build community practices, and develop improved guidelines for Scholix and Scholexplorer for easy adoption by users.

EXPERIMENTAL PROCEDURES

To collect data from Scholix API we developed Python code, which is available on UBRDA.[28] The data analysis was performed using the R statistical

programming language.[32] The code development was based on prior work from Durham University by Nicholas Syrotiuk, which was then modified to generate output based on our needs.[22] For example, the results contain many literature-to-literature links in addition to dataset-to-literature links. These could be from citing literature or links to articles in data journals. We updated the code to limit this case study to direct dataset-to-literature links. We also created a reversed version of the code that searches the API for all datasets published via a repository to identify secondary data reuse. This was not part of this analysis since the sample size is small for UBRDA.

Our code is designed to search by DOI only, since this is the main standard identifier used by DataCite and has not been tested for other forms of persistent identifiers, such as handle. In its documentation, Scholexplorer mentions that any kind of persistent identifiers can be compatible, including URLs. However, URLs are not resolved in general as the variability of the associated resolvers cannot be handled by one service.[17]

We also experienced API connection failures when testing large numbers of DOIs. To avoid disruption when collecting a large amount of metadata, the first program (get_data.py) is simpler and searches for any research output DOI that has links to at least one dataset and then creates a set of DOIs for which any dataset-literature links have been found. It also reports the total number of links found, including literature-to-literature links, the total number of dataset links found including any subsets of datasets, the total number of research outputs for which at least one unique link to a dataset has been found, and the number of research outputs for which no links were found.

The second program (metadata.py) then uses the subset of research output DOIs gathered from the first program for which one or more dataset-literature links have been found and then parses the JSON results from the API and gathers metadata for authors of research outputs, DOIs of associated datasets and dataset authors, and dataset publishers. The output is stored in tab-separated text format, which can then be cleaned and analyzed. There were some records where the cell alignments were not consistent and required manual cleaning in Excel.

For the first step we collected 31,890 research output DOIs from the University of Bath Research Portal and queried the Scholix API for links between datasets and these article DOIs, generating 1,501 results. These DOIs were then used for the second program to parse and collect associated metadata. The same data was collected twice—September 27, 2019 and November 17, 2019—to validate the consistency of the output from Scholix API.

DATA AND CODE AVAILABILITY

All data and code created during this study can be accessed from the UBRDA at https://doi.org/10.15125/BATH-00739.[28]

ACKNOWLEDGMENTS

We wish to acknowledge the University of Bath Library for funding this study.

AUTHOR CONTRIBUTIONS

N.K. wrote the codes, analyzed data, and compiled the manuscript. C.P. and M.T. reviewed and edited the manuscript and provided suggestions for improvement.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 30, 2019
Revised: January 27, 2020
Accepted: February 24, 2020
Published: April 10, 2020

REFERENCES

1. Holdren, J.P. (2013). Increasing Access to the Results of Federally Funded Scientific Research (Office of Science and Technology Policy). https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
2. Kratz, J.E., and Strasser, C. (2015). Comment: Making data count. Sci. Data *2*, https://doi.org/10.1038/sdata.2015.39.
3. Wallis, J.C., Rolando, E., and Borgman, C.L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One *8*, e67332.
4. Piwowar, H.A., Day, R.S., and Fridsma, D.B. (2007). Sharing detailed research data is associated with increased citation rate. PLoS One *2*, e308.
5. Henneken, E.A., and Accomazzi, A. (2011). Linking to data—effect on citation rates in astronomy. In Astronomical Data Analysis and Software Systems XXI. ASP Conference Series, *461*, P. Ballester, D. Egret, and N.P.F. Lorente, eds. (ASP), p. 763.
6. Drachen, T., Ellegaard, O., Larsen, A., and Dorch, S. (2016). Sharing data increases citations. Liber Quarterly *26*, https://doi.org/10.18352/lq.10149.
7. Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., Schindler, U., and Authr, C. (2017). The Scholix framework for interoperability in data-literature information exchange. D-Lib Magazine *23*.
8. Mayo, C., Vision, T.J., and Hull, E.A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. Int. J. Digit. Curation *11*, 150–155.
9. Khan, N., Thelwall, M., and Kousha, K. (2019). Data citation and reuse practice in biodiversity—challenges of adopting a standard citation model. In Proceedings of the 17th International Conference on Scientometrics & Infometrics, G. Catalano, C. Daraio, M. Gregori, H.F. Moed, and G. Ruocco, eds. (Edizioni Efesto), pp. 1220–1225.
10. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., and McGillivray, B. (2019). The citation advantage of linking publications to research data. arXiv, 1907.02565.
11. Robinson-García, N., Jiménez-Contreras, E., and Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. J. Assoc. Inf. Sci. Technol. *67*, 2964–2975.
12. Mathiak, B., and Boland, K. (2015). Challenges in matching dataset citation strings to datasets in social science. D-Lib Magazine *21*, 23–28.
13. Ghavimi, B., Mayr, P., Vahdati, S., and Lange, C. (2016). Identifying and improving dataset references in social sciences full texts. In Positioning and Power in Academic Publishing: Players, Agents and Agendas, F. Loizides and B. Schmidt, eds. (IOS Press), pp. 105–114.
14. Silvello, G. (2018). Theory and practice of data citation. J. Assoc. Inf. Sci. Technol. *69*, 6–12.
15. Scholix. (2019). Scholix: A Framework for Scholarly Link Exchange. http://www.scholix.org/home.
16. Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., and Schindler, U. (2017). The data-literature interlinking service: towards a common infrastructure for sharing data-article links. Program Electron. Libr. Inf. Syst. *51*, 75–100.
17. (2020). The OpenAIRE Scholexplorer: the data literature interlinking service. https://scholexplorer.openaire.eu/#/.
18. Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., and Simons, N. (2019). Bringing citations and usage metrics together to make data count. Data Sci. J. *18*, https://doi.org/10.5334/dsj-2019-009.
19. Hersh, G. (2019). Making open access/open data/open science a reality. Against the Grain *29*. .43. https://doi.org/10.7771/2380-176X.7782.
20. Limani, F., Latif, A., and Tochtermann, K. (2018). Linked publications and research data: use cases for digital libraries. In International Conference on Theory and Practice of Digital Libraries, E. Méndez, F. Crestani, C. Ribeiro, G. David, and J.C. Lopes, eds. (Springer), pp. 363–367.
21. Gibson, C. (2019). From Couch to Almost 5K: Raising Research Data Visibility at the University of Manchester, Library Research Plus blog https://blog.research-plus.library.manchester.ac.uk/2019/02/.
22. Syrotiuk, N. (2019). scholix, GitHub https://github.com/sefnyn/scholix.

23. Tay, A. (2018). How does Scopus find and link to related research data? Or an attempt to understand how to link datasets to articles via Scholix, Musings about librarianship blog http://musingsaboutlibrarianship.blogspot.com/2018/10/how-does-scopus-find-and-link-to.html.

24. University of Bath Research Data Archive. https://researchdata.bath.ac.uk/.

25. Gazra, K., and Fenner, M. (2018). Glad You Asked: A Snapshot of the Current State of Data Citation, DataCite Blog https://doi.org/10.5438/h16y-3d72.

26. ROR: Research Organization Registry About. https://ror.org/about/.

27. The FREYA Project The PID Graph. https://www.project-freya.eu/en/pid-graph/the-pid-graph.

28. Khan, N. (2020). Dataset for "Linking Datasets and Articles—Potentials and Challenges of Scholix Framework" (University of Bath Research Data Archive). https://doi.org/10.15125/BATH-00739.

29. Fear, K.M. (2013). Measuring and Anticipating the Impact of Data Reuse, Master's thesis (University of Michigan).

30. ORCID. ORCID: Connecting Research and Researchers. https://orcid.org/.

31. Hrynaszkiewicz, I. (2019). Publishers' responsibilities in promoting data quality and reproducibility. In Good Research Practice in Non-Clinical Pharmacology and Biomedicine Handbook of Experimental Pharmacology, *257*, A. Bespalov, M. Michel, and T. Steckler, eds. (Springer), pp. 319–348.

32. R Development Core Team (2019). R: A language and environment for statisitical computing (R Foundation for Statistical Computing).